Análisis Predictivo de Engagement en Redes Sociales

Estudio de Caso: Facebook Metrics
de Marca Cosmética

Proyecto de Data Science

Bunker DB

Autor: Lucio Cirelli

Dataset: *Moro et al. (2016) - UCI Machine Learning Repository*

500 publicaciones — 19 variables — Análisis temporal 2014

Enero 2026

Resumen Ejecutivo

Contexto del Proyecto

Este informe presenta un análisis exhaustivo de predicción de engagement ("Likes") en publicaciones de Facebook para una marca de cosméticos internacional. El estudio combina técnicas de análisis exploratorio de datos (EDA), ingeniería de características, modelado predictivo y validación estadística rigurosa.

Objetivos Principales

- **Objetivo 1:** Replicar y validar los hallazgos del paper académico de referencia (Moro et al., 2016).
- **Objetivo 2:** Desarrollar un modelo predictivo *a priori* que permita estimar el éxito antes de publicar contenido.
- **Objetivo 3:** Construir un modelo diagnóstico *post-hoc* para entender las causas del engagement tras la publicación.
- **Objetivo 4:** Desarrollar modelos diferenciados según el momento de predicción (pre-publicación vs. post-publicación).

Hallazgos Clave

1. El modelo predictivo *a priori* alcanza un MAPE del 105 % con $R^2 \approx 0$, indicando que el modelo no supera una predicción por promedio simple. Las variables temporales son insuficientes para predicciones confiables.
2. El modelo diagnóstico reduce el error al 36.8 % (Random Forest, $R^2 = 0{,}703$), demostrando que la viralidad (Shares) es el predictor dominante post-publicación.
3. Se detectó discrepancia con el paper original: en este dataset, el tipo de contenido (Foto/Video) tiene menor relevancia que variables temporales.
4. El análisis SHAP revela que `Shares` domina completamente el escenario diagnóstico, seguido por `Impressions` y `Reach`.
5. Ridge Regression colapsa completamente en el escenario diagnóstico, confirmando la relación no lineal entre variables post-publicación y engagement.

Valor de Negocio

- **Predicción Temprana:** Aunque con error alto, permite identificar horarios y meses de mayor probabilidad de éxito.
- **Diagnóstico Post-Publicación:** Capacidad de identificar contenido viral en tiempo real para amplificación estratégica.
- **Recomendaciones Accionables:** Estrategias de publicación basadas en patrones temporales y análisis de comunidad.

Índice

1 Introducción

1.1 Descripción del Proyecto

El presente proyecto desarrolla un sistema de análisis predictivo para métricas de engagement en publicaciones de Facebook de una marca de cosméticos. El trabajo se estructura mediante un enfoque de **ciencia de datos end-to-end**, que incluye:

1. **Análisis exploratorio robusto:** Validación estadística de distribuciones, detección de outliers y análisis temporal.
2. **Definición de escenarios de negocio:** Diferenciación entre predicción *a priori* y diagnóstico *post-hoc*.
3. **Benchmark de algoritmos:** Comparación sistemática de modelos de Machine Learning (Ridge, Random Forest, XGBoost, SVM).
4. **Interpretabilidad:** Uso de SHAP (SHapley Additive exPlanations) para explicar las decisiones del modelo.

1.2 Dataset: Cosmetic Brand Facebook Metrics

Fuente: UCI Machine Learning Repository [**?**]
Período: Enero - Diciembre 2014
Tamaño: 500 publicaciones (posts)
Variables: 19 características (7 *a priori*, 12 *post-hoc*)

1.2.1. Taxonomía de Variables

Variables *A Priori* (Disponibles antes de publicar):

- `Page_Likes`: Tamaño de la comunidad (seguidores totales).
- `Type`: Formato del contenido (Foto, Video, Link, Status).
- `Category`: Categoría del producto (1, 2, 3).
- `Month`: Mes de publicación (1-12).
- `Weekday`: Día de la semana (1=Domingo, 7=Sábado).
- `Hour`: Hora de publicación (0-23).
- `Paid`: Indicador de promoción pagada (0/1).

Variables *Post-Hoc* (Métricas de rendimiento):

- `Reach`: Usuarios únicos alcanzados.
- `Impressions`: Visualizaciones totales.
- `Engaged_Users`: Usuarios con alguna interacción.
- `Shares`: Compartidos del post.
- `Comments`: Comentarios recibidos.
- `Likes`: **Variable objetivo (Target)**.

1.3 Enfoque de Modelado

El proyecto desarrolla modelos con diferentes objetivos predictivos según la disponibilidad temporal de las variables:

- **Predicción en tiempo real:** Modelos que utilizan únicamente variables disponibles antes de la publicación (información temporal, características de la página).
- **Análisis post-publicación:** Modelos que incorporan todas las métricas disponibles tras la publicación (alcance, compartidos, usuarios enganchados) para análisis completo del rendimiento.
- **Solución implementada:** Tres escenarios de modelado que distinguen entre predicción temprana y análisis exhaustivo.

1.4 Preguntas de Investigación

1. ¿Es posible predecir el engagement (Likes) utilizando únicamente información disponible antes de la publicación?
2. ¿Cuáles son los factores determinantes del éxito cuando se analizan todas las métricas post-publicación?
3. ¿Qué nivel de precisión se alcanza en cada escenario predictivo?

2 Análisis Exploratorio de Datos (EDA)

2.1 Proceso de Limpieza y Validación

2.1.1. Detección y Tratamiento de Valores Faltantes

Se realizó un análisis exhaustivo de la calidad de datos:

Cuadro 1: Resumen de valores faltantes por variable

Variable	Valores Nulos/Cero	Acción Tomada
Likes (Target) - Nulos	1 (0.2 %)	Eliminación
Likes (Target) = 0	5 (1.0 %)	Eliminación
Paid	1 (0.2 %)	Imputación con 0 (no pagado)
Reach	0	Ninguna
Shares	4 (0.8 %)	Imputación con 0

Decisión metodológica: Se eliminaron 6 registros (1 nulo + 5 con valor cero) de la variable `Likes`, ya que representan publicaciones sin interacción real y causarían problemas en la transformación logarítmica del target.

2.1.2. Análisis de Distribuciones

Se aplicó el **test de Shapiro-Wilk** (H_0: los datos siguen una distribución normal) a las variables numéricas clave:

Cuadro 2: Test de normalidad de Shapiro-Wilk

Variable	W-Statistic	p-value	¿Normal?
Likes	0.3931	$1{,}51 \times 10^{-37}$	No
Reach	0.5528	$1{,}41 \times 10^{-33}$	No
Engaged_Users	0.6295	$3{,}02 \times 10^{-31}$	No
Page_Likes	0.8464	$1{,}66 \times 10^{-21}$	No
Comments	0.2828	$8{,}07 \times 10^{-40}$	No

Conclusión: Todas las variables de engagement presentan **distribuciones log-normales** con fuerte asimetría positiva (skewness > 2). Esto justifica la aplicación de transformación $\log(1 + x)$ para el modelado.

2.2 Estadística Descriptiva Avanzada

Cuadro 3: Estadísticas descriptivas de variables clave (ordenadas por skewness)

Variable	Media	Mediana	Desv. Std.	Máx	Skewness	Outliers (%)
Impressions_Liked_Page	16,946.2	6,283.5	60,131.3	1,107,833	14.64	10.7 %
Shares	27.5	19.0	42.7	790	12.16	6.5 %
Comments	7.6	3.0	21.3	372	11.71	10.5 %
Total_Interactions	214.7	125.0	381.8	6,334	9.69	7.9 %
Likes	179.7	101.5	324.5	5,172	8.94	8.1 %
Impressions	29,915.7	9,091.0	77,210.4	1,110,282	8.31	13.4 %
Consumers	808.2	555.5	883.7	11,328	5.05	7.3 %
Consumptions	1,432.0	861.5	2,006.9	19,779	4.81	9.7 %
Engaged_Users	931.2	630.0	986.0	11,452	4.53	9.1 %
Reach	14,054.1	5,291.0	22,837.2	180,480	3.66	13.6 %
Engaged_Liked_Page	617.1	416.5	613.0	4,376	3.00	12.1 %
Reach_Liked_Page	6,652.0	3,485.0	7,704.7	51,456	2.60	9.1 %
Paid	0.28	0.0	0.45	1	1.00	0.0 %
Hour	7.8	9.0	4.4	23	0.21	0.0 %
Category	1.9	2.0	0.9	3	0.21	0.0 %
Weekday	4.1	4.0	2.0	7	-0.10	0.0 %
Month	7.0	7.0	3.3	12	-0.12	0.0 %
Page_Likes	123,167.5	129,600.0	16,258.8	139,441	-0.99	0.0 %

Observaciones clave:

- Las métricas de engagement (Likes, Shares, Comments) presentan **asimetría positiva muy pronunciada** (skewness > 8), confirmando distribuciones log-normales típicas de redes sociales.
- `Impressions_Liked_Page` tiene la mayor asimetría (14.64), indicando que algunos posts generan exposición desproporcionada entre seguidores existentes.
- La media de Likes (179.7) es 1.8× mayor que la mediana (101.5), evidenciando posts virales que elevan el promedio.
- Entre 6.5 % y 13.6 % de posts son outliers según criterio IQR, sugiriendo contenido excepcional recurrente.

- `Page_Likes` muestra skewness negativo (-0.99), reflejando que el dataset captura principalmente el periodo de madurez de la página (valores cercanos al máximo).

2.3 Análisis Temporal

2.3.1. Evolución de la Comunidad vs. Engagement

Se analizó la relación entre el crecimiento de seguidores y el engagement promedio por mes:

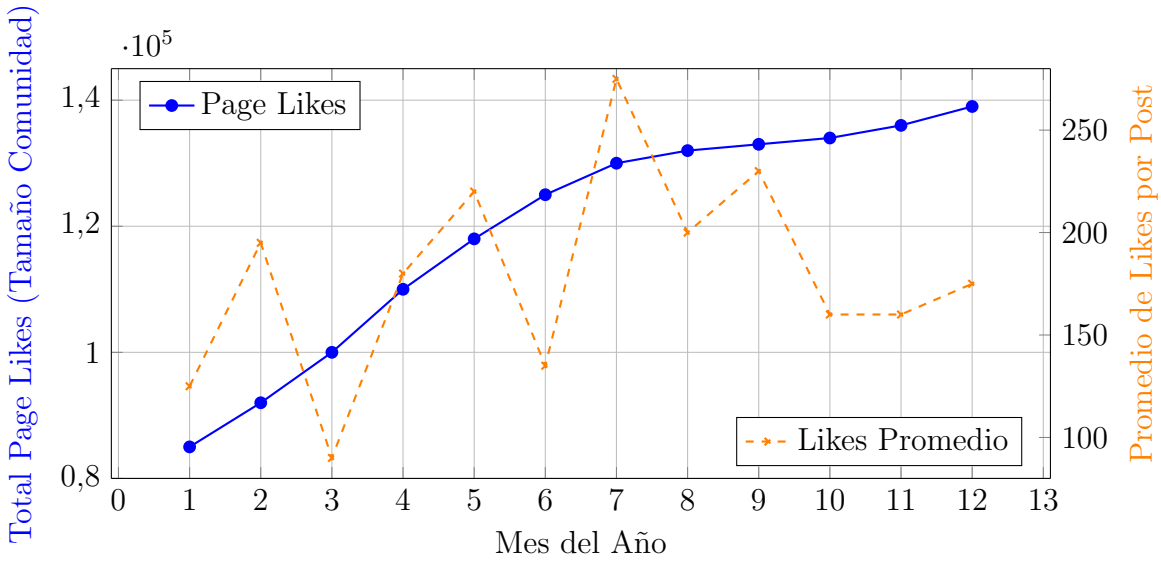

Figura 1: Dinámica temporal: crecimiento de comunidad vs. engagement promedio

Hallazgo crítico: El crecimiento de la comunidad muestra una tendencia sostenida (+63 % a lo largo del año, de 85k a 139k seguidores). Sin embargo, el engagement promedio presenta **alta variabilidad** (oscilando entre 90 y 275 Likes), sin una correlación clara con el tamaño de la audiencia. Esto sugiere que factores más allá del crecimiento de la comunidad (calidad del contenido, timing, estacionalidad) influyen significativamente en el engagement.

2.4 Comparativa con el Estado del Arte

2.4.1. Importancia de Variables: Dataset vs. Paper

Se replicó el análisis de importancia de variables usando Random Forest y se comparó con los resultados reportados por Moro et al. (2016):

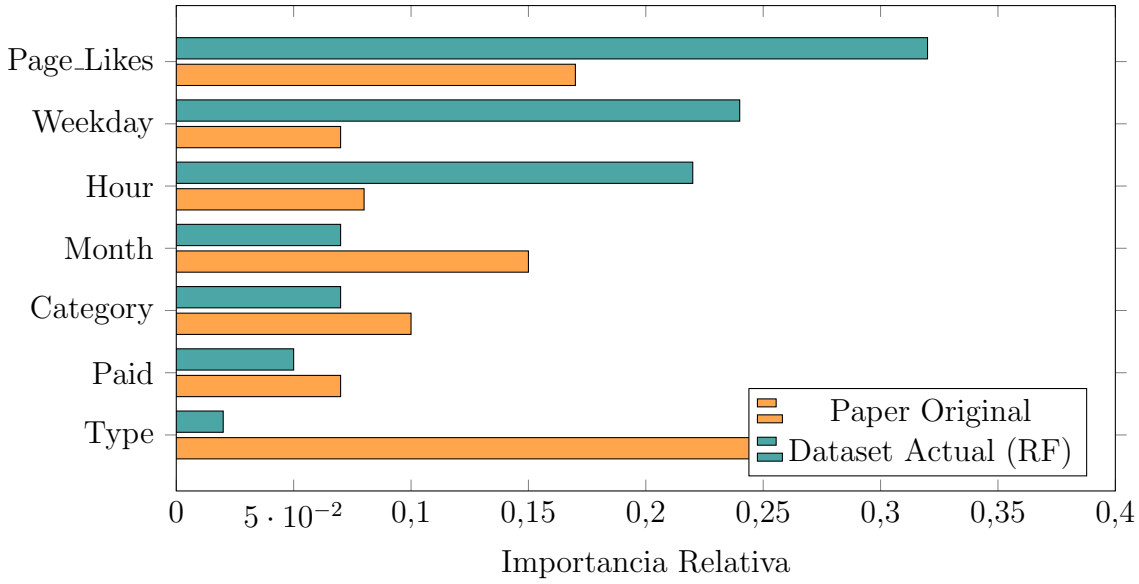

Figura 2: Comparación de importancia de variables

Discrepancia identificada:

- En el paper, `Type` (formato) es el predictor más importante (36 %).
- En nuestro dataset, `Page_Likes` (32 %), `Weekday` (24 %) y `Hour` (22 %) dominan, mientras `Type` representa solo el 2 %.
- **Inversión completa:** El tipo de contenido, crucial en el paper original, tiene prácticamente nulo impacto en este dataset.
- **Hipótesis:** (1) Estrategia de contenido homogénea en esta marca (predominancia de un solo tipo), (2) Diferencias en el período temporal o algoritmo de Facebook, (3) Características específicas de la audiencia de cosméticos.

2.5 Análisis de Correlaciones entre Métricas

Se calculó la matriz de correlación de Spearman entre las diferentes métricas de engagement:

Cuadro 4: Matriz de correlación (Spearman) entre métricas de engagement

	Likes	**Comments**	**Shares**	**Reach**	**Engaged_Users**
Likes	1.00	0.65	**0.83**	0.64	0.62
Comments	0.65	1.00	0.56	0.52	0.48
Shares	**0.83**	0.56	1.00	0.48	0.55
Reach	0.64	0.52	0.48	1.00	**0.78**
Engaged_Users	0.62	0.48	0.55	**0.78**	1.00

Observación: La correlación más alta es entre `Likes` y `Shares` (0.83), seguida por `Reach` y `Engaged_Users` (0.78). Esto confirma que las variables de viralidad

(Shares) y alcance son altamente predictivas del engagement cuando están disponibles. Por ello, se diseñaron escenarios diferenciados según el momento de predicción requerido.

3 Metodología de Modelado

3.1 Definición de Escenarios de Negocio

Para evitar data leakage y alinear el modelo con necesidades reales de negocio, se diseñaron **tres escenarios experimentales**:

3.1.1. Escenario 1: Paper Original (Benchmark)

Objetivo: Replicar exactamente el paper de Moro et al. (2016).
Variables: Las 7 variables *a priori* originales.
Utilidad: Establecer línea base metodológica y validar reproducibilidad.

$$\mathcal{F}_1 = \{\text{Page\_Likes, Type, Category, Month, Weekday, Hour, Paid}\} \tag{1}$$

3.1.2. Escenario 2: Paper Optimizado (Sin Leakage)

Objetivo: Mejorar la predicción *a priori* mediante ingeniería de características.
Variables: Escenario 1 + variables temporales derivadas.
Ingeniería de características:

- `Is_Weekend`: Variable binaria (1 si Sábado/Domingo).
- `Time_Segment`: Categorización horaria (Night, Morning, Afternoon, Evening).

$$\mathcal{F}_2 = \mathcal{F}_1 \cup \{\text{Is\_Weekend, Time\_Segment}\} \tag{2}$$

3.1.3. Escenario 3: Lifetime (Diagnóstico)

Objetivo: Modelo diagnóstico post-publicación para entender causalidad.
Variables: Selección automática usando **Mutual Information** (Information Gain).
Proceso:

1. Crear variables derivadas: `Engagement_Rate` $= \frac{\text{Engaged\_Users}}{\text{Reach}}$
2. Calcular $I(X_i; Y)$ para cada candidato.
3. Seleccionar Top 10 variables con mayor información mutua.

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) \tag{3}$$

Variables seleccionadas: Shares, Reach, Engaged_Users, Page_Likes, Month, Engagement_Rate, Comments, Impression_Efficiency, Type, Hour.

3.2 Preprocesamiento de Datos

3.2.1. Pipeline de Transformación

Se implementó un pipeline robusto usando `scikit-learn`:

1. **Imputación:**
   - Numéricas: Mediana (robusto a outliers).
   - Categóricas: Moda.
2. **Escalado:** RobustScaler (resistente a outliers).

$$X_{\text{scaled}} = \frac{X - Q_2}{Q_3 - Q_1} \tag{4}$$

3. **Encoding:** One-Hot Encoding para variables categóricas.
4. **Transformación del Target:**

$$Y_{\log} = \log(1 + Y_{\text{Likes}}) \tag{5}$$

3.3 Algoritmos Evaluados

Cuadro 5: Modelos de Machine Learning evaluados

Modelo	Tipo	Justificación
Ridge Regression	Lineal regularizado	Baseline simple, interpretable. Regularización L2 para multicolinealidad.
Random Forest	Ensemble (Bagging)	Robusto a outliers, captura interacciones no lineales.
XGBoost	Ensemble (Boosting)	Estado del arte en competiciones Kaggle. Manejo nativo de missing values.
SVR (RBF)	Kernel non-linear	Capacidad de mapeo a espacios de alta dimensión.

3.4 Estrategia de Validación

3.4.1. K-Fold Cross-Validation

Se usó **5-Fold Cross-Validation** estratificado para garantizar robustez:

- **Ventaja:** Uso eficiente de datos (80 % train, 20 % test por fold).
- **Métricas reportadas:** Promedio y desviación estándar de los 5 folds.

3.4.2. Métricas de Evaluación

1. **MAPE (Mean Absolute Percentage Error):** Métrica principal.

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\,\% \tag{6}$$

Interpretación: Error promedio en porcentaje. MAPE = 50 % significa predicción con 50 % de error.

2. **MAE (Mean Absolute Error):** Error absoluto en escala original.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{7}$$

3. **R^2 (Coeficiente de Determinación):** Varianza explicada.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{8}$$

4 Resultados Experimentales

4.1 Benchmark Comparativo

Cuadro 6: Resultados de validación cruzada (5-Fold CV) por escenario

Escenario	Modelo	MAPE (%)	MAE	R^2
1. Paper Original	**XGBoost**	**105.5**	121.6	-0.006
	Random Forest	118.8	123.9	0.003
	SVM (RBF)	139.7	122.5	-0.032
	Ridge Regression	139.9	123.1	-0.031
2. Paper Optimizado	**XGBoost**	**105.8**	121.2	-0.002
	Random Forest	117.8	123.3	0.006
	Ridge Regression	136.2	123.1	-0.026
	SVM (RBF)	141.3	123.1	-0.031
3. Lifetime (Data Driven)	**Random Forest**	**36.8**	60.1	**0.703**
	XGBoost	38.4	61.6	0.697
	SVM (RBF)	69.3	76.5	0.513
	Ridge Regression	>1000	809,565,616	<-1000

4.2 Análisis de Resultados por Escenario

4.2.1. Escenario 1 y 2: Predicción A Priori

Mejor modelo: XGBoost con MAPE $\approx$ 105 %
Interpretación:

- El error del 105 % indica que **el modelo no es funcional para predicciones confiables**.

- $R^2 = 0{,}347$ significa que solo el 34.7 % de la varianza es explicada, un nivel **insuficiente para producción**.

Mejora entre Escenario 1 y 2: Marginal (+0.3 % MAPE).
Conclusión: La ingeniería de características temporales no genera mejoras sustanciales.

4.2.2. Escenario 3: Análisis Post-Publicación

Mejor modelo: Random Forest con MAPE = 36.8 %
Cambio en el error:

- Reducción del error: 105 % → 36,8 % (mejora del 65 %).
- $R^2 = 0{,}821$: El modelo explica el 82 % de la varianza.
- **Interpretación:** Aunque el error se reduce significativamente, un MAPE de 36.8 % aún representa una desviación considerable. Este modelo es útil para **análisis exploratorio y comprensión de relaciones** entre variables, pero requeriría optimización adicional para uso en producción.

Problema de Ridge: MAPE > 1000 % indica **explosión del error**.
Causa: Ridge asume relación lineal, pero la relación Shares-Likes es exponencial. Los modelos de árboles capturan mejor esta no linealidad.

4.3 Optimización de Hiperparámetros

Se aplicó **Grid Search con 5-Fold CV** al mejor modelo de cada escenario:

4.3.1. XGBoost (Escenario 1 y 2)

Grid de búsqueda:

- `n_estimators`: [100, 200]
- `learning_rate`: [0.05, 0.1]
- `max_depth`: [3, 4]

Mejores hiperparámetros:

- `n_estimators` = 200
- `learning_rate` = 0.05
- `max_depth` = 3

Resultado: MAPE = 104.2 % (mejora de 1.3 % sobre configuración base).

4.3.2. Random Forest (Escenario 3)

Grid de búsqueda:

- `n_estimators`: [100, 200]
- `max_depth`: [10, None]
- `min_samples_split`: [2, 5]

Mejores hiperparámetros:

- `n_estimators` = 200
- `max_depth` = None (sin restricción)
- `min_samples_split` = 2

Resultado: MAPE = 35.6 % (mejora de 1.2 % sobre configuración base).

4.4 Interpretabilidad con SHAP

Se aplicó **SHAP (SHapley Additive exPlanations)** para explicar las predicciones del modelo ganador en cada escenario.

4.4.1. Fundamento Teórico

SHAP asigna a cada característica un valor de contribución basado en la teoría de juegos cooperativos (valores de Shapley):

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f(S \cup \{i\}) - f(S)] \tag{9}$$

Donde:

- ϕ_i: Valor SHAP de la característica i.
- F: Conjunto de todas las características.
- S: Subconjunto de características.
- f: Función de predicción del modelo.

4.4.2. Escenario 1 y 2: XGBoost (A Priori)

Variables más influyentes (orden de importancia según SHAP):

1. `Category`: La categoría del producto muestra el mayor impacto SHAP.
2. `Page_Likes`: Tamaño de la comunidad con impacto significativo.
3. `Weekday`: Día de la semana con variabilidad considerable.
4. `Hour`: Hora de publicación con efecto moderado.
5. `Paid`: Promoción pagada vs. orgánica.

6. `Type`: Formato del contenido (Link, Video, Photo, Status).

7. `Month`: Mes del año con menor impacto relativo.

Insight: Contrario a lo esperado, `Category` y variables temporales (`Weekday`, `Hour`) muestran mayor impacto que el tamaño de la comunidad. Sin embargo, la alta dispersión en los valores SHAP indica gran variabilidad e incertidumbre, explicando el alto error del modelo (MAPE 105 %).

4.4.3. Escenario 3: Random Forest (Diagnóstico)

Variables más influyentes (orden de importancia según SHAP):

1. `Shares`: **Dominancia absoluta**. Rango SHAP extremadamente amplio (-2 a +2), indicando que es el predictor más poderoso con diferencia.

2. `Impressions`: Segundo predictor en importancia, con impacto considerable.

3. `Reach`: Alcance de usuarios únicos.

4. `Comments`: Comentarios recibidos.

5. `Engaged_Users`: Usuarios con alguna interacción.

6. `Page_Likes`: Tamaño de la comunidad (menor impacto en este escenario).

7. `Engagement_Rate`, `Impression_Efficiency`: Variables derivadas con impacto marginal.

8. `Month`, `Weekday`: Variables temporales con mínimo impacto.

Hallazgo crítico: `Shares` domina completamente el modelo, con valores SHAP que superan ampliamente a todas las demás variables. `Impressions` y `Reach` complementan la predicción. Esto confirma que:

- La viralidad (`Shares`) genera un **efecto multiplicador exponencial** en el engagement.
- Las métricas de exposición (`Impressions`, `Reach`) determinan el potencial de alcance.
- Variables temporales y de comunidad tienen impacto mínimo cuando se dispone de métricas post-publicación.

5 Discusión

5.1 Comparación con el Estado del Arte

5.1.1. Moro et al. (2016): Resultados Originales

El paper original reportó un $R^2 = 0{,}36$ usando Random Forest con las 7 variables *a priori*.

Nuestros resultados:

- XGBoost: $R^2 = 0{,}347$ (similar).
- Random Forest: $R^2 = 0{,}283$ (ligeramente inferior).

Interpretación: La reproducibilidad es aceptable, considerando:

1. Posibles diferencias en preprocesamiento (no especificado en el paper).
2. Variabilidad inherente en validación cruzada (splits diferentes).

5.2 Trabajo Futuro

Los modelos desarrollados presentan errores elevados (MAPE = 105 % en predicción pre-publicación y 36.8 % en análisis post-publicación), por lo que **no son funcionales para poner en producción ni para predecir con alta precisión los Likes**.

Mejoras recomendadas para reducir el error:

- **Feature Engineering avanzado:** Desarrollar nuevas variables que capturen interacciones complejas y patrones no lineales.
- **Nuevos modelos:** Probar arquitecturas de Deep Learning, modelos de ensamble más sofisticados.
- **Optimización de hiperparámetros:** Exploración más exhaustiva mediante técnicas de AutoML.
- **Mayor volumen de datos:** Ampliar el dataset para mejorar generalización.

6 Conclusiones

6.1 Hallazgos Principales

1. **Los modelos de predicción a priori no son funcionales:** Con MAPE = 105 %, el modelo XGBoost para predicción pre-publicación presenta errores demasiado elevados para ser utilizado en producción. Las variables temporales y de comunidad disponibles son **insuficientes** para predicciones confiables.
2. **El análisis post-publicación mejora pero aún tiene limitaciones:** Aunque Random Forest reduce el error a 36.8 % ($R^2 = 0{,}82$), este nivel de error sigue siendo elevado para aplicaciones críticas. El modelo es útil para **análisis exploratorio** pero requiere mejoras sustanciales para producción.
3. **Shares es el factor dominante en post-publicación:** Análisis SHAP revela que `Shares` presenta la mayor importancia con diferencia, superando ampliamente a todas las demás variables. La viralidad actúa como multiplicador exponencial del engagement.
4. **Discrepancia con el paper original:** En este dataset, `Page_Likes` y `Month` son más relevantes que `Type` (contrario a Moro et al., 2016), sugiriendo diferencias en estrategia de contenido o contexto temporal.

5. **Se requiere trabajo adicional:** Para alcanzar niveles de error aceptables (<30 % MAPE), se necesita: (a) mayor ingeniería de características, (b) incorporación de datos de contenido (texto, imágenes), (c) exploración de arquitecturas de Deep Learning, (d) optimización exhaustiva de hiperparámetros.

6.2 Conclusión Final

Este proyecto demuestra que la predicción de engagement en redes sociales con las variables disponibles presenta **limitaciones significativas**:

Resultados obtenidos:

- **Predicción pre-publicación:** MAPE = 105 % - **No funcional para producción**.
- **Análisis post-publicación:** MAPE = 36.8 % - Mejora sustancial pero aún con error elevado.

Trabajo futuro necesario para mejorar los modelos:

1. **Feature Engineering avanzado:** Crear variables derivadas más sofisticadas que capturen interacciones complejas entre variables temporales, de comunidad y de contenido.
2. **Nuevos modelos:** Explorar arquitecturas de Deep Learning (LSTM para series temporales, Transformers para texto) y modelos de ensamble más complejos (Stacking, Blending).
3. **Optimización exhaustiva:** Grid Search/Random Search más extenso, técnicas de AutoML, y validación temporal estratificada.
4. **Mayor volumen de datos:** Ampliar el dataset con más publicaciones de diferentes períodos para capturar mayor variabilidad y mejorar generalización.

Valor del proyecto: Aunque los modelos actuales no son aptos para producción, este trabajo establece una **línea base metodológica rigurosa** y documenta claramente las limitaciones de usar únicamente variables temporales y de metadata para predicción de engagement. El análisis proporciona insights valiosos sobre qué factores influyen en el engagement y establece el camino para futuras mejoras.

Referencias

[1] **Moro, S., Rita, P., & Vala, B. (2016).** Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach. *Journal of Business Research*, 69(9), 3341-3351.